International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.2, pp : 193 – 197 1 Feb 2014

Text-to-Speech Synthesis using Phoneme Concatenation

Mahwash Ahmed, Shibli Nisar
National University of Computer and Emerging Sciences, (NUCES-FAST)
A.K Brohi Road H 11/4 Islamabad, Pakistan
mahwash_a@yahoo.com, shibli.nisar@nu.edu.pk

Abstract- We proposed Text-To-Speech (TTS) synthesis system based on phonetic concatenation for unrestricted input text. The input text is first converted into phonetic transcription using Letter-to-Sound rules. For synthesis of a new speech, TTS system selects the recorded phoneme units (PUs) from database and modifies the duration according to the rule based on spelling using Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA). The modified PUs are then concatenated by synchronizing pitch-periods at juncture and smoothen the transitions in order to remove the audible discontinuity and spectral mismatches. The pitch of PUs is kept to original neutral sounding.

This paper describes a simple, flexible and efficient procedure to smooth the boundaries of PUs and involves much lesser use of memory spaces. The listening test resulted on the perception of discontinuities proved that the proposed method performs better than standard TD-PSOLA system and produce highly intelligible synthetic speech.

Keywords — **Text-To-Speech, Phonemes, Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA), concatenative synthesis**

1. INTRODUCTION

Text-to-speech (TTS) synthesis transforms any linguistic information stored as data or text into speech. It is widely used in audio reading devices for blind people. In the last few years however, the use of TTS technology has grown far beyond the disabled community and become a major adjunct to the rapidly growing use of digital voice storage for voice mail and voice response systems.

Concatenative TTS synthesis system has gained in popularity in recent years, due to its more natural sounding synthesized speech. It concatenates pre-recorded speech units into the word sequences according to the pronunciation dictionary or set of rules [i]. For general purpose TTS, it must be able to read unrestricted text [ii]. Thus it is desirable to have the basic speech units much smaller, like for example phonemes or diaphones, in order to be able to synthesize all possible phonetic and prosodic variation in the language with a limited database size.

The most complex attribute of speech is the determination of correct/ appropriate prosody information for the sentence. Prosodic features consist of pitch, duration, and intensity of synthesized speech [iii].There are various methods used including statistical modeling [iv], rule based modeling [v] and hybrid modeling [vi].

Even though prosody modeling has been investigated over the years, it is still difficult to automatically derive the prosody from text. The rule based approach derives a complicated set of rules to model speaker independent prosodic variations based on the spellings. Then use these rules to modify the prosody of the original speech units.

Various digital signal processing techniques are used to modify acoustic features for concatenative speech synthesis. Among these a significant improvement came with the Pitch-Synchronous Overlap-Add (PSOLA) method [vii]. Prosodic variation based on PSOLA [viii] produces the high quality synthetic speech. Especially TD-PSOLA method is the most efficient and popular method used in concatenation synthesis techniques nowadays. However TD-PSOLA has major drawback that it does not guarantee spectral smoothing at the boundaries of concatenation speech units. This may cause speech distortion when do not fit together acoustically and suffer from audible discontinuities at concatenation points.

The objective of this paper is to overcome the limitations of TD-PSOLA and reduce the audible discontinuities between concatenation points. Phonemes are considered basic speech units in this proposed system to synthesize any speech using text for limited database size. We proposed a simple, flexible and efficient technique to modify and smooth phonemes unit (PU) boundaries and match the specified prosodic characteristics.

In this paper, we mainly focus on the duration modeling of the phonemes units according to the rule based on spelling using

standard TD-PSOLA. The modified PUs are then concatenated by synchronizing pitch-periods at juncture and smoothen the transitions in order to remove the audible discontinuity and spectral mismatches. The proposed technique presents some important advantages to TD-PSOLA. The quality of synthesized speech is influenced by the continuity of acoustic features (like pitch, amplitude etc.) at the concatenation points. The pitch of PUs is kept to original neutral sounding.

This paper is organized as follows: Section 2 discusses the proposed phonetic based TTS concatenation synthesis by concatenated by synchronizing pitch-periods at juncture and smoothen the transitions. The experimental evaluation and results are presented in Section 3; moving to the conclusive remarks in Section 4.

2. PROPOSED SOLUTION/ BASIC FRAMEWORK

The basic framework of general purpose TTS synthesis system consists of four main stages: First analyzed input text and convert to phonetic transcription. Then, modify phonemes duration based on context. Finally, synthesized speech using smoothed PUs concatenation. A simplified block diagram of our proposed solution is shown in Fig. 1.

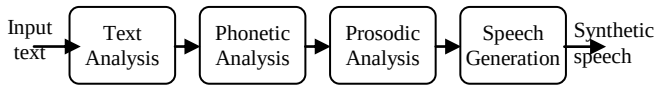

Fig. 1. Block diagram of text-speech synthesis

2.1. Text Analysis

The input text is first transformed into sentences, separates the words and expands the numerals, abbreviations and acronyms. The pre-processed text is then converted into phonetic transcription to find the correct pronunciation.

In this paper, Letter-To-Sound (LTS) rules [ix] are used for the automatic determination of the phonetic transcription of the incoming text. These rules treat the text/word as an unstructured sequence of letters and decision is made based on simple characters comparison stored in a dictionary.

There are 329 LTS rules which are very language dependent. A scanning window is passed left-to-right across each word to identify the appropriate phone sequence (41 phonemes) from the database. The basic form of a rule is as follows:

LEFT [MATCH] RIGHT = OUT

This states the particular letters MATCH appears in the specified LEFT and RIGHT context are converted into the phonemes OUT context. Any of LEFT, RIGHT or OUT context may be empty. The MATCH is searched through the database until the particular contexts are found. The last rule is context-independent default for the MATCH letters, which is converted into OUT context in the case when no other, more specific rule applies.

Thus the word ‘hello’ is represented by 4 phonemes ‘HH’, ‘AH’, ‘L’ and ‘OW’.

2.2. Duration Modeling

We mainly focus on duration modelling of each phone as it is an essential aspect of prosody and is more important for intelligibility than any other acoustic features [x]. It depends on several factors mainly the adjacent phonemes and characteristics peculiar to the phonemes.

After phonetic transcription, each phoneme unit stored in database is assigned the average duration using lookup table. The average duration statistics are similar to the one used in CMU pronunciation dictionary. Then a set of duration rules is applied to predict changes of segmental duration based on context. The phonemes may be stretched or contracted by pre-specified percentage (standard deviation) attached to each rule type as specified.

The list of rules as follows shows some of the factors taken into account, as they influence the intrinsic duration of the PUs

i) Vowel and following consonant (VC.), just before a pause in sentence, is lengthened.
ii) Vowels preceded by voiceless plosive consonant are lengthened.
iii) Postvocalic voiceless consonant shortened the duration of vowel .E.g. vowel phonemes in "ate -EY1 T" and "soup --sUW1 P" are shortened because the next consonants are voiceless.
iv) A brief pause is inserted before each sentence and at other boundaries delimited by an orthographic comma

The most commonly used technique for prosody modification in concatenative synthesis is TD-PSOLA [vii]. It enables to modify the pitch and duration of the speech independently and has low computational cost. The first step of TD-PSOLA is to determine pitch and generate pitch marks through overlapping windowed speech. For voiced PUs, we use auto-correlation function (ACF) to calculate pitch contour of each frame. We also use median filter to remove pitch errors and

ensure continuity in pitch contour. Then the pitch marks are generated by finding the maximum for each period of phone in the overlapping segments. This method can be summarized as follows:

i) For first frame, we find the maximum value of amplitude. The corresponding time (t_m) is marked as first pitch mark where m is an index for short-time (ST) signal.
ii) Search other pitch marks by finding the max in the region [tm + f * T_o, tm + (2 - f) * T_o], where T_o is the pitch period and f is factor whose range could be 0.5-0.9, usually set to 0.7[xi]. Repeat same procedure until all pitch marks are found.

For unvoiced PUs, pitch marks are regularly spaced and modified duration of particular parts of the signal. Thus Pitch-marks are placed at a pitch-synchronous rate on voiced portions of the PUs signal as shown in fig 2, and at a constant rate on unvoiced parts as in fig 3.

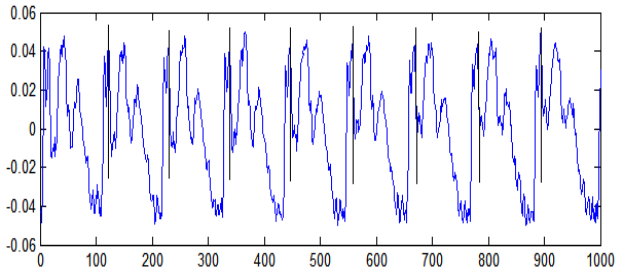

Fig.2. Pitch Marks for Voiced PU

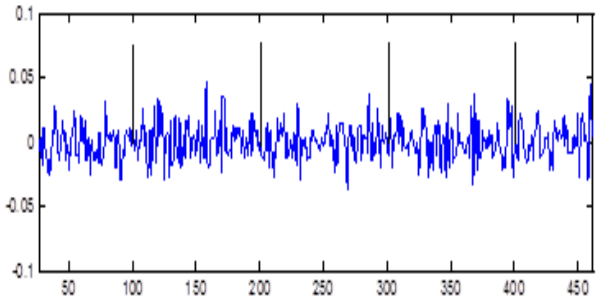

Fig 3. Pitch Marks for Unvoiced PU

Then, in order to change the duration, each pitch-marked frame is multiplied by a pitch-synchronous window hm(n) of length proportional to the local pitch period.

$$xm(n) = hm(tm- n)\, x(n) \quad (1)$$

Figure 4 shows the Hanning windows of length $2T_0$ centered on the successive pitch-marks tm and overlapped with the adjacent ST signals.

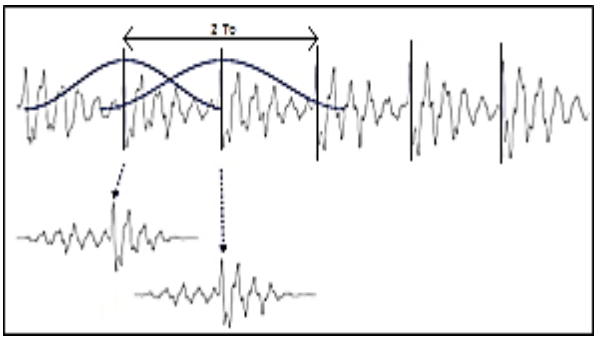

Fig. 4. Hanning Widowed Pitch synchronous frames

Next, duration modifications are applied to ST segments representation. The segments are repeated multiple times to increase duration and/or are eliminated to decrease duration. Finally, modified segments (of length $2T_0$) are added to produce the final synthetic signal using OLA. Figure 5 and 6 results in the increase and decrease of duration of vowel phoneme respectively. We use 50 % overlap to keep the final pitch contour similar to original pitch for normal speech. The greater overlap results in increase pitch and smaller overlap causes decrease pitch.

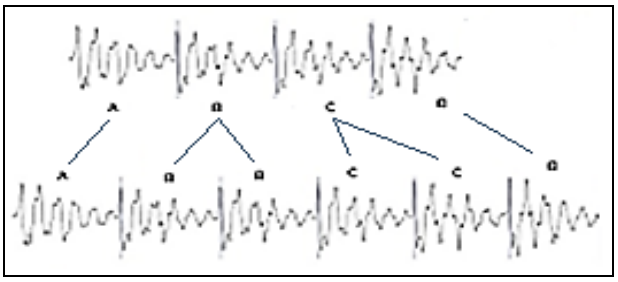

Fig.5. Duplicate ST segments to increase duration

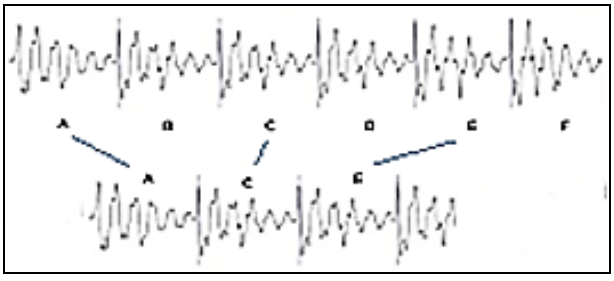

Fig.6. Remove ST segments to decrease duration

2.3. Smoothing and Speech Synthesis

Phonemes, though only 41 in number, require correct knowledge and procedures for concatenation to produce intelligible units. Ignoring co-articulation between Pus and simply concatenating

the modified duration PUs does not guarantee a good spectral smoothing at the boundaries, and thus leave artifacts which make the synthetic speech sound unnatural. The proposed method aims to remove the discontinuities and perceived click at the junctures by synchronizing pitch-marks at juncture (if both PUs are voiced) so that pitch-periods at the end of a previous PU must line up with the pitch-periods at the beginning of the current PU. A smoothing is performed on the transitional frames between consecutive PUs in order to ensure the pitch transits smoothly and minimize the spectral discontinuities at the boundary. This method can be summarized as follows:

i) Smoothly modifies the energy levels at the beginning and at the end of phonemes, by applying window to two phones to be joined so that their amplitude is close to zero at the boundary.

ii) Use overlap-add technique to concatenate the last frame of previous PU and the first frame of the current PU using equation (2) as

$$\xi_m = \boldsymbol{\alpha}\xi_p + \boldsymbol{\beta}\,\xi_c \qquad (2)$$

where ξ_p is last frame of previous PU and ξ_c is first frame of current PU. β is defined as a first n/2+1 samples of hamming window of length and α= 1- β.

Figure 7 shows the effect of smoothing and OLA of consecutive PUs. 'A' represents the last two frames of previous PU and 'B' is the first two frames of current PU. The results of applying window to both adjacent PUs are shown in 'C'. 'D' gives the results of OLA at the PUs boundaries.

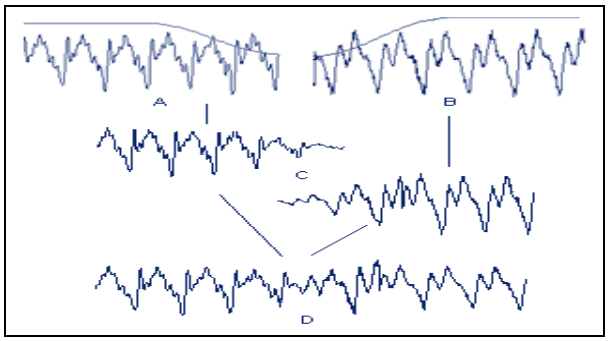

Fig. 7. Smoothing and OLA at PUs boundaries

3. EVALUATION

When evaluating PU duration modification method, two areas are of special interest: The ability of the method to remove audible discontinuities, and the resulting overall speech quality and intelligibility. A subjective test was conducted, comparing concatenation of PUs without duration modification by modified Pus using standard TD-PSOLA approach and by smoothing TD-PSOLA approach.

Phonetic speech database consists of only one unit of each 41 phoneme, pronounced with a constant (flat) pitch and same intensity for normal speaking rate. A test composed of 14 sentences, was listened by 12 – not the speech experts. The results show that the proposed TTS synthesizer produces more intelligible speech in comparison with the standard TD-PSOLA.

Spectrograms of vowel PUs synthesize using simple concatenation at boundary and proposed smoothed method as shown in fig. 8 and fig. 9.The concatenated phonemes are 'AA', 'EY' and 'OY' sampled at 11025 kHz. The proposed method greatly minimizes the spectral discontinuities at the boundary and thus results in less audible discontinuity.

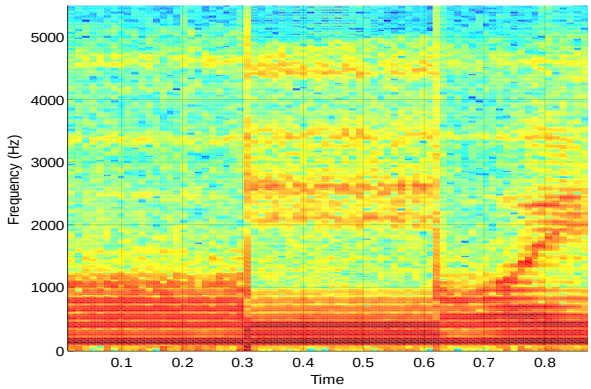

Fig.8. Spectrogram of synthesized speech using dummy concatenation

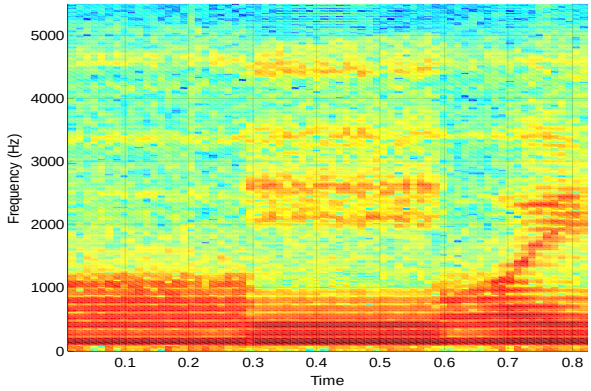

Fig.9. Spectrogram of synthesized speech using proposed smoothing concatenation

4. CONCLUSION

The results of perceptual evaluation test indicate that the synthetic speech is quite understandable. The algorithm effectively changes the duration using TD-PSOLA. The

durational modelling is crucial for naturalness. The great care is taken when modifying the duration of PUs including the precision of pitch-marks. The PUs are then smoothed at the boundary to reduce the spectral mismatches and audible discontinuities at concatenative points. This increases the quality and intelligibility of the speech and reduces the robot-like sound without an excessive increase of the computational load.

5. REFERENCES

i. *Dutoit T. "High quality Text-to-Speech Synthesis: An overview", IEEE Journal: Special Issue on Speech Recognition and Synthesize, vol. 17(5), pp. 25-37, 1997.*

ii. *Allen J. "Synthesis of speech from unrestricted text." IEEE Journal, vol.64, pp. 432-42, 1976*

iii. *Romportl J. and Kala J. "Prosody Modeling in Czech Text-to-Speech Synthesis", International Workshop on Speech Synthesis, 2007*.

iv. *Zhang W., Gu L. and Gao Y., "Recent improvements of probability based prosody model for unit selection in concatenative Text to Speech" ICASSP, pp. 3777-3780, 2009.*

v. *Buza O., Toderean G., and Domokos J., "A rule based approach to build a Text to speech system for Romanian", International Conference on Communications, pp. 33-36, 2010.*

vi. *Hung-Yan GU, Ming-Yen LAI and Sung-Feng TSAI "Combining HMM spectra models and ANN prosody models for speech synthesis of syllable prominent languages", IJASLP, pp. 451-454, 2010.*

vii. *Moulines E. and Charpentier F., "Pitch synchronous waveform processing techniques for Text-to-Speech synthesis", Speech Communication, vol 9, pp. 453-467, 1990.*

viii. *AceroA., "Source-filter models for Time-Scale Pitch-Scale Modification of speech". IEEE ICASSP, pp. 881-884, 1998.*

ix. *El'ovitz H.S., Johnson R.W., McHugh A. and John E.S. "Automatic translation of English Text to Phonetics by means of Letter-to-Sound rules" NRL Report 7948, 1976.*

x. *Carlson R. and Granstr¨om B., "A search for durational rules in a real-speech database." Phonetica, vol. 43, pp.140–154, 1986.*

xi. *Moulines, E., Hamon, C. & Charpentier, F., "High-quality prosodic modifications of speech using Time-Domain Overlap-add Synthesis." Proceedings XII GRETSI, 1989.*

xii. *Kawai H. and Tsuzaki M. (2002) "Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis.", ICSLP, pp. 2621-2624, 2002.*

xiii. *Laprie Y. and Colotte, V. "Automatic pitch marking for speech transformations via TDPSOLA". In IX European Signal Processing Conference, 1998.*

xiv. *Obin N., Xavier R. and Dujour A.L., "A multi-level context-dependent prosodic model applied to duration modeling", Inerspeech, 2009*.